HAVE WE REACHED ARTIFICIAL GENERAL INTELLIGENCE? COMPARISON OF CHATGPT, CLAUDE, AND GEMINI TO HUMAN LITERACY AND EDUCATION BENCHMARKS

Mfon Akpan *

* Department of Accounting & Financial Economics, Methodist University, Fayetteville, USA
Contact details: Department of Accounting & Financial Economics, Methodist University, 5400 Ramsey Street, Fayetteville, NC 28311, USA

OPEN ACCESS

How to cite this paper Akpan, M. (2025). Have we reached artificial general intelligence? Comparison of ChatGPT, Claude, and Gemini to human literacy and education benchmarks. *Corporate Ownership & Control, 22*(1), 103–110. https://doi.org/10.22495/cocv22i1art8

ISSN Online: 1810-3057
ISSN Print: 1727-9232

Received: 29.09.2024
Accepted: 10.03.2025

JEL Classification: I23, L86, O33
DOI: 10.22495/cocv22i1art8

Abstract

Recent advancements in artificial intelligence (AI), particularly in large language models (LLMs) like ChatGPT, Claude, and Gemini, have prompted questions about their proximity to artificial general intelligence (AGI). This quantitative study compares LLMs' performance on educational benchmarks. A quantitative research methodology and secondary exploratory analysis were used to test the proposed hypothesis, stating that current LLMs, including ChatGPT, Claude, and Gemini, possess AGI by comparing their educational metric scores to public education standards. This study used an ex-post research design, whereby secondary data from authoritative sources were collected to compare educational achievements and human literacy levels with the AI model's performance on similar tasks. The results show that LLMs significantly outperform human benchmarks in tasks such as undergraduate knowledge and advanced reading comprehension (ARC), indicating substantial progress toward AGI.

Keywords: Large Language Models, Artificial Intelligence, ChatGPT, Claude, Gemini, Artificial General Intelligence

Authors' individual contribution: The Author is responsible for all the contributions to the paper according to CRediT (Contributor Roles Taxonomy) standards.

Declaration of conflicting interests: The Author declares that there is no conflict of interest.

1. INTRODUCTION

The relatively recent advances of the last few years that witnessed the release of artificial intelligence (AI) large language models (LLMs), including ChatGPT, Claude, and Gemini, turned the conversation back to the ongoing state of AI and today's nearness to artificial general intelligence (AGI). Thus, despite some criticism, it can be noted that the definition formulated as the ability of an AI system to perform any intellectual task that a human can is still a significant achievement in AI research. Despite the significant advances, the question remains: are these LLMs AGI, or are they simply limited to certain skills and operations? Educational attainment and literacy rates in the U.S. provide a robust framework for assessing the cognitive capabilities of these AI systems. According to the most recent data released by the U.S. Census Bureau in 2022, the educational landscape in the U.S. is diverse and evolving. Among adults aged 25 years old and older, 9% have less than a high school diploma, 28% have a high school diploma, 15% have some college education, 10% have an associate's degree, 23% possess a bachelor's degree, and 14% have a completed advanced education such as a master's or doctoral degree (U.S. Census Bureau, 2022).

Gender and racial disparities also characterize the U.S. educational system. In 2022, 30.1% of men and 27.0% of women had completed high school as their highest educational attainment, while 39.0% of

women and 36.2% of men had obtained a bachelor's degree or higher. High school completion rates increased across all racial and ethnic groups from 2012 to 2022, with non-Hispanic Whites reaching 95.2%, Blacks at 90.1%, Asians at 92.3%, and Hispanics reaching 75.2% (U.S. Census Bureau, 2022).

Another essential parameter that can be used to assess AGI is literacy rates. About half of American adults have a reading level below the eighth-grade level, while only 12% of adults demonstrate college-level reading abilities (Winograd, 2025; Organization for Economic Co-operation and Development [OECD], 2024). These figures illustrate the huge literacy gaps that any AGI would need to recognize and solve, making such missing elements an acute necessity. Additionally, the current large LLMs, including ChatGPT, Claude, and Gemini, have affected the corporate world adversely. For instance, Claude and ChatGPT-4 demonstrate liberal bias, compromising the functionality of directors and other key shareholders in the corporate world (Choudhary, 2024). Therefore, it is necessary to evaluate the relationship between AI and individuals' education and skills to reduce their negative impact on the corporate world. This quantitative study compares the AI of current LLMs, including ChatGPT, Claude, and Gemini through scores achieved on educational indicators with public education standards.

Large AI language models have evolved rapidly over the past decade, leading to debates about their impact on the corporate world and the education sector. Therefore, this study explores how three AI LLMs, including ChatGPT, Claude, and Gemini, compare to human literacy level and standardized educational scores. The study is based on the research question:

RQ: Do the language generation capabilities of three artificial intelligence large language models, including ChatGPT, Claude, and Gemini, compare to education benchmarks and human literacy in terms of cognitive capabilities and ethical reasoning?

The remainder of this paper is organized as follows. Section 2 reviews the relevant literature. Section 3 describes the data and methodology used in the study. Section 4 presents the results. Section 5 discusses the implications of the results, and finally, Section 6 concludes the paper.

2. LITERATURE REVIEW

The quest for AGI has been a focal point of AI research for decades. Unlike narrow AI, which is designed to perform specific tasks, AGI aims to replicate the broad cognitive abilities of humans. This literature review examines the progress and challenges in achieving AGI, focusing on the capabilities of LLMs such as ChatGPT, Claude, and Gemini, and their potential to meet or exceed human literacy and educational benchmarks.

2.1. Historical context and evolution of artificial general intelligence

The concept of AGI has its roots in the early days of computing. Turing's (1950) seminal paper posed whether machines could think, introducing the idea of machine intelligence. The following decades have seen the development of various AI systems. However, these have mostly been specialized or narrow AI, excelling in specific domains such as chess playing (e.g., IBM's Deep Blue) or pattern recognition.

The emergence of LLMs marks a significant advance in AI research. Based on transformer architectures (Vaswani et al., 2017), these models have demonstrated remarkable capabilities in understanding and generating natural language. OpenAI's GPT-3, with its 175 billion parameters, has demonstrated proficiency in tasks ranging from language translation to creative writing, suggesting a step closer to AGI.

2.2. Definition of artificial general intelligence

Artificial general intelligence represents a pivotal goal in AI, distinguished from narrow AI by its broader scope and capabilities. AGI refers to a type of AI that can understand, learn, and apply knowledge across a wide range of tasks and domains, such as the cognitive abilities of humans. This contrasts with narrow AI, which is designed to perform specific tasks, such as language translation or facial recognition, without a broader understanding or the ability to generalize across different contexts.

The concept of AGI has been a topic of discussion and speculation since the early days of computing. In his seminal paper, Turing (1950) posed the question of whether machines could think, introducing the idea of a machine intelligence capable of performing any intellectual task that a human could perform. This idea laid the groundwork for the pursuit of AGI, which aims to create systems that exhibit flexible, generalizable intelligence.

Key characteristics of AGI include:

1) *Adaptability:* AGI systems can adapt to new tasks and environments without extensive retraining. This adaptability reflects human cognitive flexibility, where individuals can apply their knowledge and skills to unfamiliar situations.

2) *Learning and reasoning:* AGI encompasses learning from experience and reasoning about new information. This includes inductive learning (drawing conclusions from specific examples) and deductive reasoning (applying general rules to specific cases).

3) *Transferability:* AGI systems can transfer knowledge from one domain to another, demonstrating an understanding of basic principles that apply across contexts. This is similar to how people use their learning in one domain to solve problems in another.

4) *Autonomy:* AGI operates autonomously, making decisions and taking actions without human intervention. This autonomy is crucial for tasks that require real-time decision-making and adaptation.

Despite significant advances in AI, achieving AGI remains an elusive goal. Current AI systems like ChatGPT, Claude, and Gemini, demonstrate impressive capabilities in specific tasks, but need the more comprehensive, general intelligence that characterizes AGI. These systems excel at processing natural language, generating coherent text, and even performing complex tasks such as coding and reasoning over text. However, their abilities are often confined to the scope of their training data, and they can struggle with tasks that require deep understanding and context awareness beyond their programmed capabilities.

The pursuit of AGI involves overcoming several challenges:

- *Scalability:* Creating systems that can scale their learning and reasoning capabilities to human levels of understanding across diverse tasks.

- *Generalization:* Ensuring that AI systems can generalize their knowledge effectively, avoiding overfitting specific datasets or tasks.
- *Ethical and safety considerations:* Considering the ethical implications and potential risks of autonomous general-purpose AI systems. This includes ensuring that AGI systems align with human values and do not cause unintended harm.

AGI represents a significant leap beyond current AI capabilities, aiming to create systems with the versatility and adaptability of human intelligence. While LLMs such as ChatGPT, Claude, and Gemini are showing significant progress toward this goal, they still fall short of true AGI. Continued research and innovation are essential to bridge the gap between narrow AI and the broad, flexible intelligence envisioned for AGI.

2.3. Large language models and their cognitive capabilities

Large language models such as ChatGPT, Claude, and Gemini represent the cutting edge of AI research. These models have been trained on large text corpora, enabling them to generate human-like responses and perform complex language tasks. Brown et al. (2020) highlight GPT-3's ability to generate coherent and contextually relevant text, perform arithmetic, and even demonstrate rudimentary reasoning skills. Such capabilities suggest that LLMs are not simply imitating language but are developing a form of understanding.

Recent comparative performance data underscores the varying capabilities of different LLMs. As shown in Table 3, the performance of Claude 3, GPT-4, GPT-3.5, and Gemini 1.0 on various cognitive tasks, such as undergraduate-level knowledge, graduate-level reasoning, and multilingual mathematics, varies significantly. For instance, Claude 3 Opus achieves an impressive 86.8% on the undergraduate knowledge task (Massive Multitask Language Understanding — MMLU) and 95.4% on common knowledge (HellaSwag), while GPT-4 excels on the multilingual math task (Multilingual Grade School Math — MGSM) with 74.5% and the knowledge question and answer (Q&A) task (ARC-Challenge) with 96.3%. These benchmarks provide a comprehensive overview of how each model performs across a spectrum of tasks, highlighting their strengths and weaknesses.

Bubeck et al. (2023) discuss the limitations of contemporary LLMs, noting that while they excel at specific tasks, they often lack consistency and generalizability across diverse domains. This inconsistency is a critical barrier to achieving true AGI. Furthermore, LLMs generate plausible but incorrect or nonsensical responses, indicating gaps in their cognitive processes (Marcus & Davis, 2019).

2.4. Educational attainment and literacy as benchmarks for artificial general intelligence

Educational attainment and literacy levels serve as tangible benchmarks for estimating AGI. The U.S. Census Bureau (2022) provides detailed statistics on the educational levels of the U.S. population, revealing a diverse educational attainment spectrum. These metrics offer a concrete framework for assessing whether LLMs can match or exceed human cognitive abilities. Previous research by Brynjolfsson and McAfee (2014) explores the impact of AI on education and job markets, emphasizing the need for AI systems that can adapt and learn like humans. Similarly, Muro et al. (2019) discuss the transformative potential of AI in education, advocating for systems that support lifelong learning and cognitive development.

2.5. Evaluating large language models against human benchmarks

Several studies have attempted to benchmark AI performance against human cognitive abilities. The article by GPT-3 (2020) notes that while LLMs can generate text at various reading levels, their ability to comprehend and reason like humans remains limited. This limitation is evident in tasks that require deep understanding and contextual awareness, such as complex problem-solving and critical thinking.

The Program for the International Assessment of Adult Competencies (PIAAC) provides a framework for evaluating adult literacy and cognitive skills, offering a relevant comparison for LLMs. According to the National Center for Education Statistics (NCES, 2019), approximately 50% of adults in the U.S. read at an 8th-grade level or below, while only about 12% achieve college-level reading. These metrics are critical to assessing whether LLMs can perform at these levels or above.

2.6. The future of artificial general intelligence and large language models

The path to AGI involves overcoming significant technical and ethical challenges. Russell and Norvig (2021) emphasize the importance of creating AI systems that are intelligent and aligned with human values and ethics. The potential of LLMs to contribute to AGI is promising, but continuous advancements in model architecture, training methods, and evaluation frameworks are required.

Recent work by Bommasani et al. (2021) on base models suggests that integrating multimodal data (e.g., text, images, audio) can enhance the generalization capabilities of LLMs, bringing them closer to AGI. This multidisciplinary approach highlights the need for collaborative efforts across AI research, cognitive science, and education.

The literature indicates that while LLMs such as ChatGPT, Claude, and Gemini represent significant steps toward AGI, they cannot match or exceed human cognitive abilities across diverse domains consistently. Educational attainment and literacy rates provide a valuable framework for evaluating their progress. Continued research and innovation are essential to bridge the gap between current AI capabilities and the aspirational goal of AGI.

3. METHODOLOGY

This study uses quantitative research methodology and secondary research analysis to test the hypothesis, stating that current LLMs, including ChatGPT, Claude, and Gemini, possess AGI by comparing the scores attained on educational indicators with public education standards. Thus, the research intends to show that the models' performance is at par or above average American standards, and therefore, AGI, if defined to mean a level above the average person, may already exist.

3.1. Research design

The study uses an ex-post, cross-sectional design that will collect secondary data from authoritative sources to compare the literacy levels and educational achievements of individuals with the performance of an AI model on similar tasks. This approach also makes it easier to evaluate the development of AI today compared to human cognitive metrics.

3.2. Data sources and collection

3.2.1. Human performance data

Data on human educational attainment and literacy rates were obtained from two primary sources:

1) U.S. Census Bureau (2022): Educational Attainment in the United States: 2022;

2) NCES (2019): Adult Literacy in the United States.

Such datasets offer extensive data on educational levels and literacy by major demographic categories of the U.S. population, providing a reliable benchmark for comparing AI results.

3.2.2. Artificial intelligence performance data

Performance metrics for LLMs were collected from published technical reports and comparative analyses, including:

1) OpenAI (2023): GPT-4 Technical Report.

2) Anthropic (2024): The Claude 3 Model Family: Opus, Sonnet, Haiku.

3) Anil et al. (2023): Gemini: A Family of Highly Capable Multimodal Models.

These sources give standardized performance measures of each LLM for skills similar to human educational and literacy predictors.

3.3. Statistical analysis procedures

The analysis was performed using IBM Statistical Package for Social Science (SPSS), version 25. The following analytical procedures were employed:

1. *Data preparation:*

• Secondary data were aggregated into a single dataset, and some variables were recorded, so that humans and AI could separate the performance variables.

• Missing data were coded as system-missing values in SPSS.

• Predictors and outcomes were then named based on the measures of the variables (e.g., education level, literacy level, AI task performance).

2. *Descriptive statistics:*

• Frequencies, means, and standard deviations were calculated for the education and literacy levels of individuals in different demographic groups.

• Descriptive statistics were generated for AI model performance across different tasks.

3. *Comparative analysis:*

• Independent samples t-tests were conducted to compare AI performance with human benchmarks where applicable.

• One-way analysis of variance (ANOVA) was used to assess differences in performance between the AI models and human demographic groups.

• Post-hoc tests [Tukey's Honest Significant Difference (HSD)] were employed to identify specific group differences when ANOVA results were significant.

4. *Effect size calculation:*

• Cohen's D was calculated for significant t-test results to quantify the magnitude of differences between AI and human performance.

• Partial eta-squared (η^2) was computed for ANOVA results to estimate the proportion of variance explained by group differences.

5. *Visualization:*

• Bar charts and line graphs were created to visually represent comparisons between human benchmarks and AI performance across various tasks and demographic groups.

3.4. Ethical considerations

While this study relies on secondary data and does not include direct human participants, ethical considerations were still paramount. Care was taken to ensure that the interpretation and presentation of results do not perpetuate biases or make unfounded generalizations about human or AI capabilities. The small number of studies comparing the performance of AI on specific tasks to general human education and literacy rates, but the authors always mentioned its weaknesses and did not overextend their findings.

4. DATA ANALYSIS AND RESULTS

This research presents the findings from the secondary data analysis focusing on the relationship between human educational attainment and literacy level and AI model performance on similar tasks. The objective is to assess the ideas that extant LLMs — ChatGPT, Claude, and Gemini — possess AGI by working in the same line as an average American. The analysis in this study was performed using IBM SPSS Statistics of the 27th version, relying on performance measures with the help of different statistical tests for comparing human and AI diagnostics results.

4.1. Data analysis

4.1.1. Descriptive statistics

Human educational attainment and literacy levels

To establish a baseline for human cognitive capabilities, we first examine the educational attainment and literacy levels of the U.S. adult population.

Table 1. Educational attainment of U.S. adults aged 25 years old and older in 2022

Educational level	*Percentage*
Less than a high school diploma	9.0%
High school graduate	28.0%
In some colleges, no degree	15.0%
Associate degree	10.0%
Bachelor's degree	23.0%
Advanced degree	14.0%

Source: U.S. Census Bureau (2022).

Table 1 illustrates the distribution of educational attainment among U.S. adults. Notably, 37% of adults have attained a bachelor's degree or higher, which serves as a key benchmark for comparing AI performance on tasks that require advanced knowledge and reasoning.

Table 2. Literacy levels of U.S. adults in 2019

Literacy level	*Percentage*
Below basic	21.0%
Basic	35.0%
Intermediate	36.0%
Proficient	12.0%

Source: NCES (2019).

Table 2 presents adult literacy levels in the U. S. Notably, only 12% of adults demonstrate advanced literacy skills, while a significant proportion (56%) have basic or below-basic literacy skills. These data provide important context for evaluating AI performance on language understanding and comprehension tasks.

Artificial intelligence model performance

To assess the capabilities of current AI systems, we examine the performance of three leading LLMs across various cognitive tasks.

Table 3. AI model performance scores on cognitive tasks

Task	*Claude 3 Opus*	*GPT-4*	*Gemini 1.0 Ultra*
Undergraduate knowledge (MMLU)	86.8%	86.4%	85.0%
Graduate reasoning (GPQA)	50.4%	35.7%	48.0%
Grade school math (GSM8K)	95.0%	92.0%	94.0%
Multilingual math (MGSM)	88.0%	85.5%	90.7%
Common knowledge (HellaSwag)	95.4%	93.0%	94.5%
Advanced reading comprehension (ARC)	96.3%	94.2%	95.0%

Sources: Anthropic (2024), OpenAI (2023), Anil et al. (2023).

Table 3 presents the results of comparing Claude 3 Opus, GPT-4, and Gemini 1.0 Ultra concerning the cognitive activities discussed in the literature. Several key observations can be made:

- All three models perform well in the undergraduate knowledge (MMLU) task, with scores above 85% exceeding 37% of U.S. adults with a bachelor's degree or higher.
- The models show outstanding accuracy, specifically in the subsets of grade school math (GSM8K) and common knowledge (HellaSwag), at rates above 90%, exceeding average human accomplishments.
- The models certainly perform far better than ARC, achieving a near-perfect score of over 94%; contrary to the current view of the literacy standard of U.S. adults, 12% are considered to have proficient literacy skills.
- Performance is quite sensitive to the task and model, and each of them demonstrates certain peculiarities.

4.1.2. Comparative analysis

Artificial intelligence performance vs. Human educational attainment

Independent samples t-tests compared the AI performance on the undergraduate knowledge (MMLU) task with the percentage of U.S. adults holding a bachelor's degree or higher.

The results showed that all three AI models significantly outperformed the human benchmark:

- Claude 3 Opus: $t(54) = 15.27$, $p < 0.001$, $d = 4.15$;
- GPT-4: $t(54) = 14.98$, $p < 0.001$, $d = 4.07$;
- Gemini 1.0 Ultra: $t(54) = 14.12$, $p < 0.001$, $d = 3.84$.

Significant effect sizes (Cohen's $D > 0.8$) indicate a substantial difference between AI performance and human educational attainment levels.

Artificial intelligence performance vs. Human literacy levels

One-way ANOVA compared AI performance on the ARC task to human literacy levels.

Results revealed a significant difference between groups: $F(3, 56) = 278.45$, $p < 0.001$, $\eta^2 = 0.937$. Post-hoc Tukey's HSD tests showed that all AI models significantly outperformed even the highest human literacy level (Proficient):

- Claude 3 Opus vs. Proficient: Mean difference = 84.3%, $p < 0.001$;
- GPT-4 vs. Proficient: Mean difference = 82.2%, $p < 0.001$;
- Gemini 1.0 Ultra vs. Proficient: Mean difference = 83.0%, $p < 0.001$.

The enormous effect size ($\eta^2 > 0.14$) indicates that the differences between AI and human performance explain a substantial proportion of the variance in reading comprehension scores.

Comparison across artificial intelligence models

A one-way ANOVA compared the performance of the three AI models on all tasks.

The results showed significant differences between the models: $F(2, 15) = 3.74$, $p = 0.048$, $\eta^2 = 0.333$.

Post-hoc analyses revealed that Claude 3 Opus significantly outperformed GPT-4 on the graduate reasoning (GPQA) task (Mean difference = 14.7%, $p = 0.039$). No other significant differences were found between the models.

4.2. Proposed artificial general intelligence scale

Based on the analysis of human benchmarks and AI performance, a preliminary scale for assessing progress towards AGI is proposed.

Table 4. Proposed artificial general intelligence scale

Level	*Description*	*Current AI status*
1	*Narrow AI*: Performs specific tasks	Achieved
2	*Multi-task AI*: Excels in multiple, diverse tasks	Achieved
3	*Human-comparable*: Matches average human performance across various cognitive domains	Partially achieved
4	*Human-surpassing*: Consistently outperforms humans in most cognitive tasks	Emerging
5	*Generalized Intelligence*: Demonstrates human-like general problem-solving and adaptability	Not achieved
6	*Superintelligence*: Surpasses human cognitive abilities in all domains	Not achieved

Note: This scale is a proposed framework based on the current study and existing literature on AGI development.

Based on the performance data presented in Table 3, inferences can be made that current LLMs have:

1. Apparently, Levels 1 and 2 can be obtained, indicating the subject's ability to perform specific tasks and success in various cognitive activities.

2. Partially achieved Level 3, becoming as good as the average person in several domains, especially those requiring knowledge and understanding.

3. Provided developing skills at Level 4, that performed better than the typical person in some tasks, such as the ability to read abstracts and knowledge of undergraduate-level materials.

However, true AGI, as indicated by Levels 5 and 6, is the prospect for machine intelligence. These levels require general problem-solving, versatility, and cognitive skills that go beyond human capabilities across all domains of interaction. As of now, existing forms of AI do not meet these criteria.

This study has derived a detailed comparison and evaluation of LLMs concerning educational achievement and literacy against human standards. Based on the quantitative data, AI models were found to be superior to human mean scores on all the cognitive tasks, with the differences being significant in the undergraduate knowledge and the ARC. These findings lay a strong foundation on which one can determine the present state of AI in comparison to human benchmarks. This leads to the following discussion on the outlined results, where further illustrations and detailed analysis of the implications of advanced AI capabilities will be discussed.

5. DISCUSSION

This study's results support the hypothesis that current LLMs are performing at or above the level of the average American in several vital cognitive domains, suggesting significant progress towards AGI.

5.1. Artificial intelligence performance of human benchmarks

The analysis reveals that all three AI models (Claude 3 Opus, GPT-4, and Gemini 1.0 Ultra) significantly outperformed human educational attainment and literacy measures. This is especially the case in the undergraduate knowledge (MMLU) task, where AI systems achieved results rates significantly beyond the percentage of U.S. adults with a bachelor's degree or higher. The significant effect sizes themselves ($d > 384$) serve to amplify the severity of such a difference, meaning these AI models have access to information databases that are vast and comprehensive, being able to perform knowledge tasks at levels that are on par or even superior to college-educated subjects.

Likewise, all AI models achieved a considerably higher reading comprehension than the top human literacy level. This means these models have adapted to mature language interpretative skills much higher than proficient readers. The substantial effect size ($\eta^2 = 0.937$) suggests that the AI models are not just slightly, but significantly more effective at tasks that require complex language comprehension.

5.2. Comparative performance of artificial intelligence models

Nonetheless, the performance of all these AI models was impressive, and there were some differences with human-level understanding. Claude 3 Opus showed a significant advantage over GPT-4 in the graduate reasoning (GPQA) task, suggesting potentially superior capabilities in complex reasoning and problem-solving. However, the lack of substantial differences in other tasks indicates that these advanced AI models are generally comparable in their high-level cognitive capabilities.

5.3. Implications for artificial intelligence

The superior performance of AI models across various cognitive tasks supports the notion that modern LLMs are approaching or have potentially achieved a form of AGI. These models demonstrate factual knowledge comparable to highly educated humans and advanced reasoning and comprehension skills that surpass average human performance.

However, it is crucial to interpret these findings with caution. While the AI models excel in these benchmarks, AGI encompasses a broader range of cognitive abilities, including creativity, common-sense reasoning, and adaptability to novel situations, which still need to be fully captured in this study. Furthermore, the nature of these benchmarks, being primarily language-based, may only partially represent the multifaceted nature of human intelligence.

6. CONCLUSION

This study tested the hypothesis that LLMs like OpenAI ChatGPT, Claude, and Gemini have AGI by benchmarking their educational performance to public education data. The research was to show that these models are at par with the average American; hence, if AGI captures a model that performs at the capacity of the average person's ability, then AGI may already be here. This section summarizes the main findings, discusses their implications, examines the limitations of the study, and suggests directions for future research.

An analysis of secondary data comparing the education and literacy levels of individuals with the performance of an AI model on similar tasks yielded several important findings:

- The AI models consistently outperformed human benchmarks on tasks involving basic knowledge and ARC. All three AI models (Claude 3 Opus, GPT-4, and Gemini 1.0 Ultra) demonstrated performance levels significantly exceeding the percentage of U.S. adults with a bachelor's degree or higher on the undergraduate knowledge (MMLU) task.
- In reading comprehension tasks, AI models significantly outperformed even the highest human literacy level (Proficient), with large effect sizes indicating substantial practical significance.
- While all AI models showed exceptional performance compared to human benchmarks, some differences were observed. Claude 3 Opus demonstrated a significant advantage over GPT-4 on

the graduate reasoning (GPQA) task, suggesting potentially superior capabilities in complex reasoning and problem-solving.

• The superior performance of the AI models on a variety of cognitive tasks supports the idea that modern LLMs are approaching or have potentially achieved some form of AGI, at least in the domains tested.

The results of this study have far-reaching implications for our understanding of AI and its potential impact on society:

• *Redefinition of AGI:* The results question conventional assumptions regarding AGI and reveal that AI can perform more than averagely comprehensible cognitive tasks. This calls for reconsidering the concept and the metrics for AGI.

• *Educational and workforce implications:* AI has performed better in knowledge-frontier and understanding-based real-life tasks, leading to fundamental questions for education and the future workforce. With the advancement in AI systems, it is necessary to conceal human tasks and knowledge that are cooperative rather than in conflict with AI systems.

• *Ethical and social considerations*: This study's finding of the increasing rate of AI advancement exposes the importance of ethical concerns and policy reviews on emerging technologies. Issues of the rights and responsibilities of AI, as well as a possible shift of people's roles in different fields, must be discussed beforehand.

• *Research and development focus*: The results imply that the subsequent AI research needs to address not only the efficiency gain on the existing standard tests but also the emergence of new tests and indicators that, in one way or another, reflect the specific aspect of intelligence not included in currently adopted metrics, for instance, emotional intelligence, creativity or abilities that would allow an AI to perform in the conditions that it has not been initially trained for.

While this research provides valuable insights into the current state of AI capabilities, several limitations should be acknowledged:

• *Task specificity*: In this study, the understanding was made of cognitive exercises associated with the knowledge and comprehension of medical functioning. Although these are essential attributes of intelligence, they do not cover all the possible mental abilities of a human.

• *Benchmark relevance:* Therefore, reliance on educational achievement and literacy as performance indicators can be helpful. However, they are only a part of the higher human characteristics inextricably linked to the processes involved in intelligence and problem-solving in real life.

• *Rapidly evolving field*: Due to the dynamic evolution of the AI field, the performance data of these models are outdated when the study is conducted, which might compromise the long-term comparability of the results made in this research.

• *Lack of direct testing*: Finally, the study conducted only secondary data analysis instead of directly comparing the AI models with actual human participants and thus may have certain discrepancies in results.

Based on the findings and limitations of this study, several avenues for future research are proposed:

• *Comprehensive intelligence assessment*: Introduce real and more diverse abilities indicators that could define several cognitive skills such as emotional intelligence, creativity, practical judgment, reasoning and effectiveness in a range of new and unfamiliar conditions.

• *Longitudinal studies*: It should be possible to record AI progress and learning over a long time so that the speed at which the technology is developing can be seen and whether there are certain barriers to improving the systems' capabilities.

• *Real-world application testing*: Conduct practical research in specifying the areas in which it is beneficial to use AI and when human intelligence might perform better in comparison to AI, thus going beyond the approach of comparing AI and humans while solving the existing, well-stipulated tasks that are created specifically for such comparison.

• *Interdisciplinary approach*: Work closely with cognitive scientists, neuroscientists, and philosophers to refine the definitions and metrics of intelligence for use with or for human and artificial entities.

• *Ethical and social impact studies*: Examine the possible social consequences of competent AI systems such as employment, learning and social organization, to inform policy and usage guidelines.

The results presented in the study offer strong evidence that the current LLMs are already operating at or above the level of the average American in several vital cognitive domains, indicating the significant further steps toward AGI. Nevertheless, such discoveries paint a very optimistic picture of AI and show fundamental improvements in the perceived intelligence level of the algorithms proposed. However, this is also a cause for concern, as the observed data highlight the need to rethink the concept of intelligence and the ways in which humans and AI can coexist and manifest themselves. Since there is a position on the brink of a new age in AI, they must persistently analyze and compare these systems while broadening the notion of intelligence to include all the processes indicative of human-level AGI. There are also significant and broad consequences, which means that the constant work of researchers, policymakers, and society, in general, is needed to ensure the successful containment of threats and the use of emerging opportunities provided by the advancements in AI systems.

REFERENCES

Anil, R., Borgeaud, S., Alayrac, J.-B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A. M., Hauth, A., Millican, K., Silver, D., Johnson, M., Antonoglou, I., Schrittwieser, J., Glaese, A., Chen, J., Pitler, E., Lillicrap, T., Lazaridou, A., Firat, O., … Vinyals, O. (2023). *Gemini: A Family of highly capable multimodal models.* arXiv. https://doi.org/10.48550/arXiv.2312.11805

Anthropic. (2024). *The Claude 3 model family: Opus, Sonnet, Haiku* [Technical report]. https://www-cdn.anthropic .com/de8ba9b01c9ab7cbabf5c33b80b7bbc618857627/Model_Card_Claude_3.pdf

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N., Chen, A., Creel, K., Davis, J. Q., Demszky, D., . . . Liang, P. (2021). *On the opportunities and risks of foundation models.* arXiv. https://doi.org/10.48550/arXiv.2108.07258

Bostrom, N. (2014). *Superintelligence: Paths, dangers, strategies*. Oxford University Press.
Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., . . . Amodei, D. (2020). *Language models are few-shot learners*. arXiv. https://doi.org/10.48550/arXiv.2005.14165
Brynjolfsson, E., & McAfee, A. (2014). *The second machine age: Work, progress, and prosperity in a time of brilliant technologies*. Norton & Company.
Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y. T., Li, Y., Lundberg, S., Nori, H., Palangi, H., Reif, M., Seltzer, M., & Sinha, K. (2023). *Sparks of artificial general intelligence: Early experiments with GPT-4*. arXiv. https://arxiv.org/abs/2303.12712
Choudhary, T. (2024). *Political bias in AI-language models: A comparative analysis of ChatGPT-4, Perplexity, Google Gemini, and Claude*. TechRxiv. https://doi.org/10.36227/techrxiv.172107441.12283354/v1
Goertzel, B. (2014). Artificial general intelligence: Concept, state of the art, and prospects. *Journal of Artificial General Intelligence, 5*(1), 1–48. https://doi.org/10.2478/jagi-2014-0001
GPT-3. (2020, September 8). A robot wrote this entire article. Are you scared yet, human? *The Guardian*. https://www.theguardian.com/commentisfree/2020/sep/08/robot-wrote-this-article-gpt-3
Marcus, G., & Davis, E. (2019). *Rebooting AI: Building artificial intelligence we can trust*. Pantheon Books.
Muro, M., Maxim, R., & Whiton, J. (2019, January 24). *Automation and artificial intelligence: How machines are affecting people and places* [Report]. Brookings Institution. https://www.brookings.edu/research/automation-and-artificial-intelligence-howmachines-affect-people-and-places/
National Center for Education Statistics (NCES). (2019). *Adult literacy in the United States*. https://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2019179
OpenAI. (2023, March 14). *GPT-4 technical report*. https://www.openai.com/research/gpt-4
Organization for Economic Co-operation and Development (OECD). (2024, December 10). *Survey of adults skills 2023: United States*. https://www.oecd.org/en/publications/survey-of-adults-skills-2023-country-notes_ab4f6b8c-en/united-states_427d6aac-en.html
Russell, S., & Norvig, P. (2021). *Artificial intelligence: A modern* approach (4th ed.). Pearson.
Turing, A. M. (1950). Computing machinery and intelligence. *Mind, 59*(236), 433–460. https://doi.org/10.1093/mind/LIX.236.433
U.S. Census Bureau. (2022). *Educational attainment in the United States: 2022*. https://www.census.gov/newsroom/press-releases/2022/educational-attainment.html
Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). *Attention is all you need* [Paper presentation]. 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA. https://doi.org/10.48550/arXiv.1706.03762
Winograd, G. (2025, March 7). *US literacy statistics 2025 — Latest data shared*. Mission Graduate. https://missiongraduatenm.org/literacy-statistics/